An efficient algorithm for global alignment of protein-protein interaction networks

Đỗ Đức Đông
Vietnam National University-Hanoi
dongdoduc@vnu.edu.vn

Đặng Thanh Hải
Vietnam National University-Hanoi
hai.dang@vnu.edu.vn

Trần Ngọc Hà
Thai Nguyen University of Education
hatn84@gmail.com

Đặng Cao Cường
Vietnam National University-Hanoi
cuongdc@vnu.edu.vn

Hoàng Xuân Huấn
Vietnam National University-Hanoi
huanhx@vnu.edu.vn

***Abstract*— Global alignment of two protein-protein interaction networks is an essentially important task in bioinformatics/computational biology field of study. It is a challenging and widely studied research topic in recent years. Accurately aligned networks allow us to identify functional modules of proteins and/or orthologous proteins from which unknown functions of a protein can be inferred. We here introduce a novel efficient heuristic global network alignment algorithm called FASTAn, which includes two phases: the first to construct an initial alignment and the second to improve such alignment by exerting a repeated local optimization procedure. The experimental results demonstrated that FASTAn outperformed SPINAL, the state-of-the-art global network alignment method in terms of both commonly used objective scores and the running time.**

Keywords — FASTAn, Heuristic algorithm, Biological network alignment, Protein-protein interaction networks.

Introduction

Prior to the advent of network alignment in bioinformatics/computational biology, identification of orthologous proteins was only based on evolutionary relationship, which is often denoted by the sequence homology [1, 21]. It is, however, not adequate for identifying conserved protein complexes [10, 22, 24]. The emergence of advanced high-throughput bio-technologies over the last decade has allowed characterizing protein-protein interaction network (PPI) more accurately for various organisms. Such these networks posed a number of interesting network analysis problems [3, 6, 13-15], such as network topology analysis [8], module detection [2], etc. Among these problems, aligning networks is crucially important, which provides valuable information for prediction of protein functions or verification of known functions of proteins [7,9, 23].

PPI network alignment methods fall into two approaches: local alignment and global alignment. For the former, the objective is to identify sub-networks with similar topology and/or conserved sequence homology in aligned networks [11, 12, 19, 22]. Generally, the result of a local alignment includes many overlapped sub-networks since a protein can be aligned with multiple proteins in the other network, causing the ambiguity. The objective of the latter approach is to avoid the ambiguity as in local alignment by drawing an injection between proteins in two different networks. Global alignment of two networks was proven to be NP-hard by Aladag and Erten [1].

The first noticeable global network alignment method is IsoRank [23] proposed by Sing et al., (2008) which is based on local alignments. Afterwards, a number of similar algorithms have been developed. PATH and GA [24], PISwap [4, 5] introduced appropriate relaxation over the cost function on a set of random matrices or applied local searches over existing local alignments generated by other algorithms. MI-GRAAL [13,14] and its variants [17,18] were based on combination of greedy techniques with heuristics information such as graphlet, group classification coefficients, eccentricities and similarity value (E-value from BLAST). These algorithms are all faster in producing better results when compared with others previously proposed. They were, however, optimized only for either an objective function or scalability, but not both. Because PPI networks are very often of large node number both accuracy and scalability (in the sense of running time) are equally important. Very recently, Aladag and Erten (2013) proposed SPINAL algorithm [1], which has been demonstrated to produce the best resulting alignments fastest. SPINAL is a heuristic algorithm with polynomial time, comprising two phases: the first to calculate homology scores for every pair of proteins in two networks;

978-1-4673-8374-5/15/$31.00 ©2015 IEEE

the second to build an injection by locally improving every subset of available solutions.

This paper proposes a novel algorithm called FASTAn for global alignment of protein-protein interaction networks. The algorithm includes two phases: the first one to build an initial alignment and the second one to enhance such alignment by local optimization. Our experimental results show that FASTAn outperforms SPINAL (the state-of-the-art PPI alignment method) in term of the running time and alignment quality defined by the corresponding objective function.

The remainder of this paper is structured as follows. Section 2 present a formal concept of the network alignment problem and some associated issues. The proposed algorithm FASTAn is introduced in section 3. Section 4 then describes our experiments and the performance comparisons between FASTAn and SPINAL. Finally, conclusion and perspective works are presented afterwards.

I. Global Alignment Problem Of PPI Networks And Related Works

We denote two protein-protein interaction networks by $G_1 = (E_1, V_1)$ and $G_2 = (E_2, V_2)$, where V_1, V_2 indicate sets of nodes corresponding to proteins in the network G_1, G_2, respectively; E_1, E_2 indicate sets of edges corresponding to protein-protein interactions in G_1, G_2, respectively. Without loss of generality we can assume that $|v_1| < |v_2|$ where $|v|$ denotes the element number of V.

Network alignment aims at finding an injection from V_1 into V_2 which is the best according to specific evaluation criteria. There currently has no formally clear definition of these criteria. In the following definition we make use of criteria, which have been exerted in previous related studies [1,4,5,14, 23].

Definition 1. (Network alignment) The graph $A_{12} = (V_{12}, E_{12})$ is considered as an alignment of two network if and only if:

i. Each node $< u_i, v_j > \in V_{12}$ corresponds a pair of nodes $u_i \in V_1$ and $v_j \in V_2$.

ii. Two distinct nodes $< u_i, v_j >$ and $< u_i', v_j' >$ of V_{12} imply $u_i \neq u_i'$ and $v_j \neq v_j'$.

iii. The edge $(< u_i, v_j >, < u_i', v_j' >)$ belong to E_{12} if and only if $(u_i, u_i') \in E_1$ and $(v_j, v_j') \in E_2$.

Definition 2. (Optimal global alignment of PPI networks) An alignment $A_{12} = (V_{12}, E_{12})$ is a solution to the problem of aligning two protein network G_1, G_2 globally if it maximizes the *global network alignment score* as in the Eq. (1):

$$GNAS(A_{12}) = \alpha |E_{12}| + (1-\alpha) \sum_{\forall < u_i, v_j >} similar(u_i, v_j) \quad (1)$$

where $a \in [0,1]$ is the parameter to balance the relative importance between the network topology similarity and the sequence similarity. The value $Similar(u_i, v_j)$ is approximated using the BLAST bit-scores or E-values.

According to a study by Aladag and Erten[1], the problem of finding optimum global network alignment is NP-hard. They proposed a polynomial time algorithm called SPINAL with the complexity being:

$$SPINALComplexity = O(k \times |V_1| \times |V_2| \times \Delta_1 \times \Delta_2 \times log(\Delta_1 \times \Delta_2)) \quad (2)$$

Where k is the number of times the main loop being executed (According to [1] the algorithm converges after looping 10-15 times); Δ_1, Δ_2 are the largest node degree of the network G_1, G_2 respectively.

Their experiments on benchmark datasets of protein networks on Saccharomyces cerevisiae, Drosophila melanogaster, Caenorhabditiselegans and Homo sapiens revealed the outperformance of SPINAL over IsoRank and MI-GRAAL, which are two state-of-the-art methods by then.

II. FastAn Algorithm

A. Algorithm description

The algorithm FASTAn includes two phases: the first to build an initial alignment and the second to improve such alignment by a local optimization procedure called *Rebuild*.

Initial alignment building

Given two graph G_1, G_2, a value of the parameter α, similarity scores between node pairs $< u_i, v_j >$ of V_1, V_2, respectively and each subset of node pairs $V_{12} \subset V_1 \times V_2$, we denote

$V_{12}^1 = \{u_i \in V_1 :< u_i, v_j > \in V_{12}\}$, $V_{12}^2 = \{v_j \in V_2 :< u_i, v_j > \in V_{12}\}$

The FASTAn procedure in Fig.1 will perform the following steps:

Step 1. Initialize V_{12} with a node pair $< u_i, v_j >$ of the largest similarity score

Step 2. Loop from *k = 2* to $|V_1|$

2.1. Find a node $u_i \in V_1 - V_{12}^1$ that has the maximum number of edges connecting to nodes in V_{12}^1;

2.2. Find a node $v_j \in V_2 - V_{12}^2$ such that when adding thepair $< u_i, v_j >$ into V_{12} the *GNAS(A12)* value (see Eq. 1) gets maximal. Such node j is called the best matching node (u_i, V_{12});

2.3. Add node $< u_i, v_j >$ into V_{12};

2.4. Update E_{12} based on V_{12}

Step 3. Perform loops to improve $A_{12} = (V_{12}, E_{12})$ with the procedure *Rebuild*.

We note that, at steps 2.1 and 2.2, it is possible to have more than one node to be the best. In this case the procedure will choose a random node among such.

After building successfully an initial alignment FASTAn jumps to phase 2, in which the procedure *Rebuild* is exerted to improve the quality of such initial alignment.

Algorithm 1. Procedure of FASTAn

Input: Graph 1: $G_1 = (V_1, E_1)$; Graph 2: $G_2 = (V_2, E_2)$;
Similarities of node pairs;
Balancing parameterα
Output: Alignment network $A_{12} = (V_{12}, E_{12})$
Begin
V_{12} = {<u_i,v_j>} //The best similar pair $< u_i, v_j >$
for k=2 **to** $|v_1|$ **do**
u_i = find_next_node(G_1);
v_j = choose_best_matched_node (u_i, G_1, G_2) ;
$V_{12} = V_{12} \cup < u_i, v_j >$;
Update(E_{12});
end-for
Rebuild(A_{12});
End

Figure 1. Specification of FASTAn procedure

Rebuild procedure

Given A_{12} resulted from phase 1 and a predefined n_{keep} value (1% by default) to specify the number of nodes in the set SeedV_{12}, the procedure Rebuild in Fig.2 will perform as follows:

Step 1. Create a set $SeedV_{12}$ of V_1 comprising n_{keep} nodes in V_1 with top scores that are calculated as follows:

$$score(u) = \alpha \times w(u) + (1-\alpha) \times similar(u, f(u)) \quad (3)$$

where $u \in V_1$ and $f(u) \in V_2$ that is aligned with u in A_{12}, $w(u)$ is the number of nodes $v \in V_1$ such that $(u,v) \in E_1$ and $(f(u), f(v)) \in E_2$.

Step 2. Update V_{12} using $SeedV_{12}$ and A_{12}

Step3. Perform the loop as Step 2 of phase 1 with $k = n_{keep} + 1$ until $|V_1|$ to identify A_{12}

Algorithm 2 *Rebuild* procedure

Input:Graph 1: $G_1 = (V_1, E_1)$; Graph 2: $G_2 = (V_2, E_2)$;
Alignment network A_{12};n_{keep}
Output: Better alignment network $A_{12} = (V_{12}, E_{12})$
Begin
Build $SeedV_{12}$;
Build V_{12}; // based on $SeedV_{12}$ and G_{12}
for k=n_{keep}+1 **to** $|V_1|$ **do**
u_i = find_next_node(G_1);
v_j = choose_best_matched_node (u_i, G_1, G_2);
$V_{12} = V_{12} \cup < u_i, v_j >$;
Update(E_{12})
end-for
end.

Figure 2. Specification of the Rebuild procedure

After every execution of the procedure *Rebuild* we have a new alignment that is then taken as input A_{12} for the next Rebuild run. This is looped until no improvement of *GNAS(A_{12})* obtained.

B. FASTAn complexity

It is obvious to see that the complexity of phase 1 and each loop in phase 2 of the algorithm FASTAn is:

$$O\left(|V_1| \times (|E_1| + |E_2|)\right) \quad (4)$$

The number of times phase 2 being looped in our experiments does not exceed 20. Combining $|V_1| \times \Delta_1 \geq E_1$ and the complexity of SPINAL as defined in Eq. 2 we have:

$$|V_1| \times |V_2| \times \Delta_1 \times \Delta_2 \geq E_1 \times E_2 > \left(|V_1| \times (|E_1| + |E_2|)\right) \quad (5)$$

The complexity of FASTAn is therefore of lower order than that of the SPINAL.

III. Experiment

Experiments have been done to compare the proposed algorithm FASTAn and SPINAL (the state-of-the-art network alignment method) on 4 benchmark datasets that had been used in the study of SPINAL [1]. FASTAn were exerted with different values of n_{keep} parameter, including 1%, 5%, 10%, 20% and 50%. The experiment results showed that the n_{keep} value of 1% allows FASTAn to yield the best performance. We here therefore present the performance of FASTAn with the n_{keep} parameter of 1%. The comparison criteria are GNAS and edge correctness (EC) measures. Although we already presented the complexity comparison between two algorithms we also compared the average running time of both. The experiments were done on a PC computer with CPU Intel Core 2 Duo 2.53GHz, RAM DDR2 4GB and Ubuntu 13.10 64 bit operation system.

A. Data

We used 4 benchmark datasets that had been used to evaluate SPINAL performances by its authors [1]. They are datasets of protein-protein interactions on: Saccharomyces cerevisiae (sc), Drosophila melanogaster (dm), Caenorhabditis elegans(ce), and Homo sapiens (hs). These networks were obtained from [20]. A description of these network, including protein and interaction number, are shown in Table 1. It therefore has 6 different pair of networks (*ce-dm*, *ce-hs, ce-sc, dm-hs, dm-sc, hs-sc*) to be aligned. The parameter α gets 5 possible values, namely 0.3, 0.4, 0.5, 0.6 and 0.7 as used in [1].

TABLE 1
DESCRIPTION OF 4 BENCHMARK DATASETS OF PROTEIN-PROTEIN INTERACTIONS

Dataset	No. of proteins	No. of interactions
ce	2805	4495
dm	7518	25635
sc	5499	31261
hs	9633	34327

B. Experimental results

As alluded to in Section 3.1 that the FASTAn is a random algorithm, it was executed 100 times for each pair of study

PPI networks. The GNAS, EC and running time were averaged over those calculated from such 100 resulting alignments. They were then compared with those of SPINAL, which had been reported in [1] (See Table 2). The corresponding 95% CI of these scores of FASTAn are presented in Table 3. The comparisons of running time between FASTAn and SPINAL are shown in Table 4.

TABLE 2.

COMPARISONS OF FASTAN AND STATE-OF-THE-ART GLOBAL NETWORK ALIGNMENT ALGORITHM SPINAL ACCORDING TO GNAS AND EC CRITERIA USING DIFFERENT VALUES OF THE PARAMETER α. EACH CELL SHOWS TWO VALUES, INCLUDING THE OBJECTIVE FUNCTION'S SCORE GNAS (ABOVE) AND EC NUMBER (BELOW). THE VALUES IN BOLD INDICATE THE OUTPERFORMANCE OF FASTAN OVER SPINAL.

Datasets	α = 0.3		α = 0.4		α = 0.5		α = 0.6		α = 0.7	
	FASTAn	**SPINAL**	**FASTAn**	**SPINAL**	**FASTAn**	**SPINAL**	**FASTAn**	**SPINAL**	**FASTAn**	**SPINAL**
ce-dm	**778.46** **2560.7**	717.99 2343.0	**1034.20** **2564.6**	941.19 2320.0	**1290.11** **2567.2**	1159.93 2300.0	**1545.86** **2567.7**	1350.59 2237.0	**1801.24** **2567.6**	1586.87 2258.0
ce-hs	**863.46** **2842.8**	728.26 2370.0	**1144.17** **2838.1**	993.07 2446.0	**1429.89** **2844.9**	1229.95 2437.0	**1708.81** **2838.0**	1501.61 2487.0	**1994.87** **2843.4**	1764.93 2512.0
ce-sc	**834.79** **2761.1**	709.12 2326.0	**1109.93** **2761.2**	963.28 2384.0	**1389.21** **2769.7**	1168.95 2323.0	**1663.39** **2766.5**	1422.74 2361.0	**1936.83** **2763.1**	1683.13 2398.0
dm-hs	**2260.31** **7478.3**	1883.22 6189.0	**3007.11** **7481.9**	2517.23 6235.0	**3755.36** **7429.0**	3160.48 6282.0	**4496.45** **7478.2**	3790.79 6291.0	**5242.32** **7478.8**	4451.6 6344.0
dm-sc	**1977.82** **6569.7**	1579.06 5203.0	**2631.85** **6565.5**	2075.14 5150.0	**3290.03** **6570.7**	2668.65 5311.0	**3950.16** **6577.4**	3180.27 5283.0	**4603.41** **6572.3**	3759.07 5360.0
hs-sc	**2268.21** **7531.8**	1731.81 5703.0	**3017.96** **7528.5**	2253.66 5593.0	**3772.96** **7535.2**	2839.00 5651.0	**4520.51** **7527.0**	3434.54 5706.0	**5279.88** **7538.1**	4066.22 5798.0

TABLE 3

95% CI OF THE SCORE GNAS (ABOVE IN EACH CELL) AND EC (BELOW IN EACH CELL) OF THE PROPOSED METHOD FASTAN CALCULATED FOR EACH PAIR OF STUDIED PPI NETWORKS WITH DIFFERENT VALUES OF THE PARAMETER α.

Datasets	α = 0.3	α = 0.4	α = 0.5	α = 0.6	α = 0.7
ce-dm	776.71-780.20 2554.76-2566.71	1031.87-1036.53 2558.56-2570.55	1287.52-1292.69 2561.92-2572.38	1542.58-1549.15 2562.15-2573.19	1797.47-1805.01 2562.15-2572.97
ce-hs	861.38-865.54 2835.66-2849.91	1141.54-1146.81 2831.40-2844.80	1426.24-1433.55 2837.49-2852.23	1704.59-1713.04 2830.9-2845.1	1936.13-2014.11 2836.73-2850.15
ce-sc	832.71-836.88 2753.99-2768.20	1107.08-1112.78 2754.07-2768.39	1385.35-1393.07 2761.98-2777.5	1658.72-1668.07 2758.7-2774.36	1931.82-1941.84 2755.95-2770.31
dm-hs	2257.83-2262.8 7469.99-7486.6	3003.68-3010.53 7473.26-7490.54	3751.37-3759.36 7478.89-7494.99	4491.11-4501.78 7469.29-7487.1	5236.36-5248.29 7470.22-7487.3
dm-sc	1975.58-1980.05 6562.24-6577.18	2628.55-2635.16 6557.19-6573.79	3285.91-3294.15 6562.41-6578.91	3944.38-3955.95 6567.72-6586.99	4596.57-4610.25 6562.5116-6582.07
hs-sc	2265.05-2271.38 7521.13-7542.37	3013.83-3022.09 7518.17-7538.89	3767.3-3778.62 7523.85-7546.57	4514.5-4526.5 7516.92-7537	5272.06-5287.69 7526.93-7549.27

TABLE 4.

THE AVERAGE RUNNING TIME (IN SECOND) OF FASTAN AND THAT OF SPINAL WHEN BOTH ARE RUN TO ALIGN EACH PAIR OF STUDIED PPI NETWORKS ON THE SAME PC.

Data sets	ce-dm	dm-sc	dm-hs	ce-hs	hs-sc	ce-sc
SPINAL	540.2	1912.1	1736.8	664.3	2630.6	638.2
FASTAn	221.5	1064.5	1395.9	327.9	1507.8	142.2

Experimental results reveal that FASTAn was able to find out solutions (i.e. global alignments) having significantly higher GNAS and EC values than that of SPINAL (p-value <2.2e-16, which is calculated using t-test on GNAS and EC values of 100 resulting alignments) for all α values on 6 available network pairs. Interestingly, the worst alignments among those generated from 100 times running FASTAn on all network pairs were all better than the corresponding alignments generated by SPINAL.

IV. CONCLUTION AND FUTURE WORKS

In this article we proposed a novel algorithm called FASTAn including two phases for global alignment of two protein-protein interaction networks. The first phase builds an initial alignment while the second exerts a local optimization procedure to improve the quality of the initial alignment.

Experimental results demonstrated the advancement and efficacy of the proposed algorithm in global alignment of protein-protein interaction network in terms of GNAS, EC criteria and running time as well. The authors of SPINAL also introduced another version of SPINAL that is optimized for the Gene Ontology Coherence (GOC) measure. In the future we will develop FASTAn following this direction.

Finally, the procedure *Rebuild* of FASTAn depends on a critical parameter called n_{keep}, which is a number of nodes with top scores in the previous alignment retained after each repetition. They are considered as correctly aligned and then

used to find alignments for other nodes remaining. Therefore, setting n_{keep} to a very large value can produce bad alignments and vice versa not enough information to align well. Currently, the value of n_{keep} is empirically chosen by comparing the performance of FASTAn on a number of n_{keep} values. Although the chosen value (1%) does not guarantee to be optimum but makes sense since it allowed FASTAn to outperformance the state-of-the-art related method. We hence set 1% as the default value of the n_{keep} parameter. It is worth studying further to get the optimal value of this parameter automatically in the future.

ACKNOWLEDGMENT

This work was done during the research stay in the Vietnamese institute for advanced study in mathematics (VIASM). This work has also been partly supported by Vietnam National University, Hanoi (VNU), under Project No. QG.15.21.

REFERENCES

[1] Aladag, A.E. and Erten, C. (2013), *SPINAL: scalable protein interaction network alignment*. Bioinformatics, Vol. 29 no 7, 917–924

[2] Bader,G.D. and Hogue,C.W. (2002), *Analyzing yeast protein-protein interaction data obtained from different sources*. Nat. Biotechnol., 20, 991–997.

[3] Banks,E. et al., (2008),*NetGrep: fast network schema searches in interactomes*. Genome Biology, 9,R138

[4] Chindelevitch,L. et al. (2010), *Local optimization for global alignment of protein interaction networks*. In: Pacific Symposium on Biocomputing,Hawaii,USA, pp. 123–132

[5] Chindelevitch L. et al. (2013*), Optimizing a global alignment of protein interaction networks*, Bioinformatics ,Vol. 29 no. 21,2765–2773

[6] Dost,B. et al. (2008),*QNet: a tool for querying protein interaction networks*. J. Comput. Biol., 15, 913–925

[7] Dutkowski,J. and Tiuryn,J. (2007), *Identification of functional modules from conserved ancestral protein–protein interactions*. Bioinformatics, 23, i149–i158.

[8] Han,J.D. et al. (2004), *Evidence for dynamically organized modularity in the yeast protein-protein interaction network*. Nature, 430, 88–93.

[9] B.H. Junker and F. Schreiber, *Analysis of Bological Networks*, wiley, 2008

[10] Kelley,B.P. et al. (2003), *Conserved pathways within bacteria and yeast as revealed by global protein network alignment*. Proc. Natl Acad. Sci. USA, 100, 11394–11399.

[11] Kelley,B.P. et al. (2004*), Pathblast: a tool for alignment of protein interaction networks*. Nucleic Acids Res., 32,83–88.

[12] Koyuturk,M. et al. (2006),*Pairwise alignment of protein interaction networks*. J. Comput. Biol., 13, 182–199.

[13] Kuchaiev,O. et al. (2010), *Topological network alignment uncovers biological function and phylogeny*. J. R. Soc. Interface., 7, 1341–1354.

[14] Kuchaiev,O. and Przulj,N. (2011) *Integrative network alignment reveals large regions of global network similarity in yeast and human*. Bioinformatics, 27, 1390–1396.

[15] Kuhn HW: *The Hungarian Method for the assignment problem*. Naval Res Logistics Q 1955, 2:83-97.

[16] Liao,C.S. et al. (2009) *IsoRankN: spectral methods for global alignment of multiple protein networks*. Bioinformatics, 25, i253–i258.

[17] Memisevic,V. and Przulj,N. (2012), *C-graal: common-neighbors-based global graph alignment of biological networks*. Integr. Biol., 4, 734–743.

[18] Milenkovic,T. et al. (2010), *Optimal network alignment with graphlet degree vectors*. Cancer Inform.,Vol.9, 121–137.

[19] Narayanan,M. and Karp,R.M. (2007), *Comparing protein interaction networks via a graph match-and-split algorithm.* J. Comput. Biol., Vol. 14, 892–907.

[20] Park,D. et al. (2011) *IsoBase: a database of functionally related proteins across PPI networks*. Nucleic Acids Res., 39, 295–300

[21] Remm,M. et al. (2001), *Automatic clustering of orthologs and in-paralogs from pairwise species comparisons*. J. Mol. Biol., 314, 1041–1052.

[22] Sharan,R. et al. (2005), *Conserved patterns of protein interaction in multiple species*. Proc. Natl Acad. Sci. USA, 102, 1974–1979.

[23] Singh,R. et al. (2008), *Global alignment of multiple protein interaction networks*. In: Pacific Symposium on Biocomputing. pp. 303–314.

[24] Zaslavskiy,M. et al. (2009) *Global alignment of protein-protein interaction networks by graph matching methods*. Bioinformatics, Vol.25, 259–2.